Book Review: Data Quality: Concepts, Methodologies and Techniques by C. Batini and M. Scannapieco

Reviewed by Heather Maguire

School of Management and Marketing,
University of Southern Queensland,
Queensland, Australia
E-mail: maguireh@usq.edu.au

Published 2006 by Springer, New York, USA. pp.151–167. ISBN: 13 978-3-540-33172-8

Keywords: book review.

Reference to this book review should be made as follows: Maguire, H. (2007) 'Data Quality: Concepts, Methodologies and Techniques by C. Batini and
M. Scannapieco', Int. J. Information Quality, Vol. x, No. x, pp.xxx-xxx.

Biographical notes: Heather Maguire lectures in Management at the University of Southern Queensland. Her PhD investigating psychological contracting in the banking industry was awarded in 2001. She also holds a BEd, MBus and a MBA. She has consulted for industry in a range of areas including the impact of change in the Australian finance sector and administrative process improvement within a range of industries. Her research interests include information quality and the link between data and information quality and corporate governance. Other research areas of interest include trust, relational capital and psychological contracting, and the concept of virtual organisations.

1 Introduction

Batini and Scannapieco provide a comprehensive overview of approaches to data quality in an era during which there is increasing recognition of the cost of poor data quality, increased focus on provenance and trustworthiness of data and an absence of enforced standards on classification and definitions of data dimensions and metrics. Their publication should prove of benefit to researchers, academics, students and practitioners in the area of data and information quality.

2 Purpose

The authors set out to convince the reader of the importance of data quality to decisional and operational processes and to describe current techniques and methodologies used to address the measurement and improvement of data quality.

3 Scope

Recognition is paid to the multidisciplinary nature of the concept of data quality but the focus in this publication is on electronic data which the authors describe as "representing real world objects in a format that can be stored, retrieved and elaborated by a software procedure and communicated through a network (p.6)". The authors provide a good overview of data quality issues together with a comprehensive description of known and reliable approaches to data quality.

The book begins with an explanation of the meaning of data quality and of its multidimensional nature together with the need for more sophisticated management of data quality following the advent of wide scale use of the Internet and networked environments. The authors address the evolution of data from the structured data typical of relational databases to semi structured data and unstructured data, documents, images, sounds and maps which has caused continuous change in the concept of data quality. The authors propose that this evolution is likely to continue as ICT technology is applied to an every widening range of sciences and real world scenarios. The dimensions of data quality are then discussed with reference made in particular to the work of Wand and Wang (1996), Wang and Strong (1996) and Redman (1996). The need for a broad range of dimensions is explained on the basis of data attempting to represent all kinds of spatial, temporal and social phenomena.

Having defined data quality, its multidimensional nature and the evolution taking place within the concept, a number of models are presented for dealing with data quality dimensions. The Polygen model (Wang and Madnick, 1990) is described in the context of structured data, D^2Q (Scannapieco et al., 2004) in the context of semi structured data and IP-MAP (Shankaranarayan et al., 2000) for management of information systems. IP-UML (Scannapieco et al., 2005) as a extension of IP_MAP is also discussed. The authors propose that the future of research relating to models for dealing with data quality dimensions may rest in the areas of provenance and trustworthiness.

An overview of quality composition and error localisation is provided. This includes a useful discussion of costs and benefits of data quality models using cost classifications developed by English (1999), Eppler and Helfert (2004) and Loshin (2004). Discussion of benefits classification is also provided. The chapter provides sufficient detail to be useful in guiding the reader in adopting an appropriate data quality model based on specific usage context.

Probabilistic, empirical and knowledge-based object identification techniques are described. The impact of evolutions in networking and Internet technologies accompanied by the development of the XML standard are proposed as improving the mechanisms to represent the semantics of data. The authors provide an outline of the steps involved in object identification and provide a description of seven object identification techniques including Fellegi and Sunter (1969), Cost-based (Verykios et al., 2003), Sorted neighbourhood and variants (Stolfo and Hernandez 1995), Delphi (Ananthakrishna et al., 2002), DogmatiX (Weis and Naumann, 2005), Intelliclean (Low et al., 2001) and Atlas (Tejada et al., 2001). Following a detailed comparison of Cost-based, probabilistic and empirical techniques, the conclusion is reached that probabilistic techniques are most commonly used because of their relative maturity.

Data integration is described as a major research and business area that aims to allow users to access data stored by heterogeneous data sources through the presentation of a unified view of the data. The authors explain that data integration aims to overcome technological, schema and instance level heterogeneities. They provide an overview of proposals to deal with both quality-driven query processing and instance-level conflict resolution. Discussion is provided of proposals to perform quality-driven query processing including QP-alg (Naumann et al., 1999), DaQuinCis (Scannapieco et al., 2004), and Fusionplex. These models are compared according to a range of criteria. An overview is then provided of the following techniques proposed to solve instance-level conflicts – SQL (Naumann and Haussler, 2002), Aurora (Yan and Ozsu, 1999), Fusionplex (Motro and Ragov 1998), DaQuinCIS (Scannapieco et al., 2004), FraSQL (Schallehn et al., 2002), and OORA (Lim and Chiang 1998). An interesting theoretical perspective on inconsistencies in data integration is provided.

Description of the data quality measurement and improvement process is one of the strengths of this publication. Data Quality (DQ) methodology is defined and a comprehensive list of the types of knowledge involved in the data quality measurement and improvement process provided together with a clear mapping of the input/output structure of a general-purpose methodology for assessing and improving data quality. Models are classified as data driven vs. process driven, measurement vs. improvement, general purpose vs. special purpose and intraorganisational vs. interorganisational. The aims of assessment methodologies are discussed in terms of providing a precise evaluation and diagnosis of the state of the information system with regard to DQ issues in terms of a number of parameters including

measurements of the quality of data bases and data flows
costs to the organisation due to low quality data
comparison with data quality levels considered acceptable from experience or benchmarking together with suggestions for improvement.

The authors provide a comparative analysis of three general-purpose methodologies for data quality measurement and improvement as proposed in the literature. These methodologies include TDQM (Shankaranarayan et al., 2000), TQdM (English 1999) and Istat (Falorsi and Scannapieco, 2006). A comprehensive analysis of the basic common phases among the methodologies is provided. The authors conclude that while TDQM and TQdM may be regarded as suitable for specific organisations or information products, Istat is more suited to interorganisational analysis. TQdM is proposed as the methodology most suited to managers. The authors (p.181) propose an original methodology Complete Data Quality Management (CDQM) (Aimetti et al., 2005) which they claim provides a reasonable balance between completeness and the practical feasibility of the data quality improvement process which can be utilised in relation to all types of knowledge. Comprehensiveness, flexibility and simplicity of application are proposed as the advantages of CDQM and a useful application of the proposed model to a case study is provided.

The critical requirement for tools and frameworks in order to make techniques and methodologies effective is discussed. The authors suggest that the choice of tools should be addressed only after a full understanding has been reached of relationships between organisations, processes, databases, data flows, external sources, dimensions and activities. The authors have chosen to provide their discussion of tools based on relevant research rather than the large number of commercial tools available for data quality issues. An overview is provided of tools for both single organisations and cooperative

information systems. This discussion is supplemented with a description of toolboxes which can be utilised to compare tools. A detailed description of six tools includes

a critical evaluation
activities addressed by each tool
main features
the application domains in which usage of each tool has been reported.

Six tools suitable for use within single organisations are discussed. These include Raman and Hellerstein's Potter's wheel (Raman and Hellerstein, 2001), Caruso et al.'s Telcordia's tool (Caruso et al., 2000), Galhardas et al.'s Ajax (Galhardas et al., 2001), Vassiliadis et al.'s Artkos (Vassiliadis et al., 2001), Buechi et al.'s Choice Maker (Cluemaker)[1] (Buechi et al., 2003) and Low et al.'s Intelliclean (Low et al., 2001). The authors point out that three of these tools i.e., Potter's Wheel, Artkos and Intelliclean are academic prototypes while the other three are commercial products. Scannapieco's DaQuinCIS framework and Matra's Fusionplex framework are discussed as suitable for use within cooperative information systems. Discussion then progresses to two toolboxes for comparing tools. Firstly, Neiling et al.'s (2003) theoretical framework is described. Secondly Tailor (Elfeky et al., 2002), a toolbox for comparing object identification techniques and tools through experiments is overviewed.

The book concludes with an interesting discussion of possible future developments within the data quality research area. The

problems associated with defining a reference set of data, quality dimensions and metrics are discussed and a number of research issues are flagged as needing further investigation. These research issues (p.222) include

defining a comprehensive set of metrics allowing an objective assessment of the quality of a database
appropriate measurement methods
characterisation of quality of data in the context of information services
- possible tradeoffs within the set of dimensions characterising quality of data.

The chapter highlights three object identification issues. These include the need to focus on
description of principal research challenges characterising XML object identification
problem and research issues of object identification of information in the personal information (PIM) context
relationships between record linkage and privacy.

Evolution in data integration issues and methodologies for measuring and improving data quality are also discussed. The need to more closely relate data quality issues and business process issues at operational, tactical and strategic level is highlighted and the chapter concludes with a description of Pernici and Scannapieco's proposed model
(p.232) which is designed to improve to associate and improve quality of information to web data and a methodology for data quality design and management of web information systems.

4 Content quality

The book provides a comprehensive coverage of data quality issues and methodologies for assessment and improvement of data quality as well as a glimpse into what the future may hold for data quality issues.

5 Style

The publication is enhanced by conciseness and clarity of expression. This feature is enhanced by wide use of diagrams, figures and tables to highlight key points and to provide relevant comparison of methodologies, techniques etc. This is supplemented by appropriate use of italics and variation in font size and style to stress key points throughout. Structure is also well thought out throughout the publication. The sequence of the chapters and the internal organisation of the chapters follow a logical sequence. The referencing technique utilised has the potential to cause reader frustration on occasions e.g., p.147 "other proposals are present in the literature, including [81, 154] and p.168 in the headings for Figures 7.4 and 7.5". Utilisation of the Harvard Referencing technique may have enhanced readability.

6 Other features

The structure allows for an unusual feature of proposed usage of this publication. The authors recognise the dual nature of their potential audience and propose two possible reading paths based on a differing selection of chapters for the reader to follow. One path is considered suitable for an academic audience the other for a practitioner audience.

7 Application

The reader is provided with practical solutions, methodologies and benchmarks relating to data error localisation and correction, object identification and data integration. The content could provide relevant within a postgraduate student environment and includes a two-part complex data quality project to provide students with a glimpse of real world problems and potential solutions.

8 General comments

In the section describing inputs and outputs of a DQ measurement and improvement methodology, good comparison is made between a number of assessment methodologies and improvement solutions and the authors propose a new methodology – Complete data Quality Management (CdQM) and apply this methodology to a case study organisation. This section could prove useful for postgraduate students who could consolidate and/or

assess their knowledge of this area through replicating the application of CdQM to an organisation selected by their lecturer/supervisor or of their own choice.

9 Conclusion

Overall a well organised, comprehensive view of the current data quality issues and a glimpse of what the future may hold. The book is of potential use to a wide audience of researchers, academics and practitioners in the area of assessment and improvement in data quality.

References

Aimetti, P., Missier, P., Scannapieco, M., Bertoletti, M. and Batini, C. (2005) 'Improving government-to-business relationships through data reconciliation and process re-engineering', in Wang, R., Pierce, E., Madnick, S. and Fisher, C. (Eds.): Advances in Management Information Systems – Information Quality, Sharpe, ME, p.151.

Ananthakrishna, R., Chaudhuri, S. and Ganti, V. (2002) 'Eliminating fuzzy duplicates in data warehouses', Proceedings of VLDB, Hong Kong, pp.586–597.

Buechi, M., Borthwick, A., Winkel, A. and Goldberg, A. (2003) 'ClueMaker: a language for approximate record matching', Proceedings of 8th International Conference on Information Quality, Cambridge, pp.207–223.

Caruso, F., Cochinwala, M., Ganapathy, U., Lalk, G. and Missier, P. (2000) 'Telecordia's database reconciiation and data quality analysis tool', Demonstration at VLDB, pp.615–618.

Elfeky, M., Verykios, V. and Elmagarmid, A. (2002) 'Tailor: A record linkage toolbox', Proceedings of 18th International Conference on Data Engineering, San Jose CA, pp.17–28.

English, L. (1999) Improving Data Warehouse and Business Information Quality, Wiley and Sons, New York.

Eppler, M. and Helfert, M. (2004) 'A classification and analysis of data quality costs', Proceedings of 9th International Conference on Information Quality, Boston.

Falorsi, P. and Scannapieco, M. (2006) Principi Guida per la Qualita dei Dati Toponomastici nella Pubblica Amministrazione (in Italian), http://www.istat.it/dati/pubbsci/contributi/ Contr_anno2005.htm.

Fellegi, I. and Sunter, A. (1969) 'A theory for record linkage', Journal of the American Statistical Association, Vol. 64, No. 328, pp.1103–1210.

Galhardas, H., Florescu, D., Shasha, D., Simon, E. and Saita, C. (2001) 'Declarative data cleaning: language, model and algorithms', Proceedings of VLDB, Rome, Italy, p.312.

Lim, E. and Chiang, R. (1998) 'Global object model for accommodating instance heterogeneities', Proceedings of ER'98, Singapore, pp.435–448.

Loshin, D. (2004) Enterprise Knowledge Management – The Data Quality Approach, Morgan Kaufmann Series in Data Management Systems, USA.

Low, W. Lee, M. and Ling, T. (2001) 'A knowledge-based approach for duplicate elimination in data cleaning', Information Systems, Vol. 26, No. 8, pp.585–606.

Motro, A. and Ragov, I. (1998) 'Estimating quality of databases', Proceedings of 3rd International Conference on Flexible Query Answering Sytems (FQAS'98), Roskhilde, Denmark, pp.298–307.

Naumann, F. and Haussler, M. (2002) 'Declarative data margins with conflict resolution', Proceedings of 7th International Conference on Information Quality (IQ2002), Cambridge, pp.212–224.

Naumann, F., Leser, U. and Freytag, J (1999) 'Quality-driven Integration of Heterogeneous Information Systems', Proceedings of VLDB'99, Edinburgh, UK, pp.447–458.

Neiling, M., Jurk, S., Lenz, H. and Naumann, F. (2003) 'Object identification quality', Proceedings of DQCIS, Siena, Italy, Available at http://edoc.hu-berlin.de/oa/conferences/reZyPELZpaa4w/ PDF/27UP92RdcFjp2.pdf

Raman, V. and Hellerstein, J. (2001) 'Potter's wheel: an interactive data cleaning system', Proceedings of VLDB, Rome, Italy, pp.381–390.

Redman, T. (1996) Data Quality for the Information Age, Artech House, Boston.

Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M. and Baldoni, R. (2004) 'The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems', Information Systems, Vol. 29, No. 7, pp.551–582.

Scannapieco, M., Pernici, B. and Pierce, E. (2005) 'IP-UML: a methodology for quality improvement based on IP-MAP and UML', in Wang, R., Pierce, E., Madnick, S. and Fisher, C. (Eds.): Advances in Management Information Systems – Information Quality, pp.651–682.

Schallehn, E., Sattler, K. and Saake, G. (2002) 'Extensible and similarity-based grouping for data integration', Proceedings of ICDE, San Jose CA, p.277.

Shankaranarayan, G., Wang, R. and Ziad, M. (2000) 'Modeling the manufacture of an information product with IP-MAP', Proceedings of 5th International Conference on Information Quality, Cambridge, pp.1–16.

Stolfo, S. and Hernandez, M. (1995) 'The merge/purge problem for large databases', Proceedings of SIGMOD, San Jose CA, pp.127–138.

Tejada, S., Knoblock, C. and Minton, S. (2001) 'Learning object identification rules for information integration', Informatioin Systems, Vol. 26, No. 8, pp.607–633.

Vassiliadis, P., Vagena, Z., Skiadopoulos, S., Karayannidis, N. and Sellis, T. (2001) 'ARTKOS: toward the modelling, design, control and execution of ETL processes', Information Systems, Vol. 26, pp.537–561.

Verykios, V., Moustakides, G. and Elfeky, M. (2003) 'Bayesian decision model for cost optimal record matching', The VLDB Journal, Vol. 12, pp.28–40.

Wand, Y. and Wang, R. (1996) 'Anchoring data quality dimensions in ontological foundations', Communications of the ACM, Vol. 39, No. 11, pp.88–95.

Wang, R. and Madnick, S. (1990) 'A polygen model for heterogeneous database systems: the source tagging perspective', Proceedings of VLDB'90, Brisbane, Australia, pp.11–17.

Wang, R. and Strong, D. (1996) 'Beyond Accuracy: what data quality means to data consumers', Journal of Management Information Systems, Vol. 12, No. 4, pp.6–34.

Weis, M. and Naumann, F. (2005) 'DogmatiX tracks down duplicates in XML', Proceedings of SIGMOD, pp.431–442.

Yan, L. and Ozsu, T. (1999) 'Conflict tolerant queries in AURORA', Proceedings of CoopIS'99, Edinburgh, UK, p.279.

Note

[1] Buechi et al. refer to this technique as ClueMaker not ChoiceMaker.